The Huber-Pollard Z-theorem

Wellner; 5/13/98

The Context/Setup

Suppose that $\Theta \subset R^k$, and that

$$\begin{aligned} \Psi_n &: \Theta \to R^k\,, \quad \text{random maps} \\ \Psi &: \Theta \to R^k\,, \quad \text{deterministic maps} \end{aligned}$$

Suppose that $\widehat{\theta}_n$ and θ_0 are the corresponding solutions (or approximate solutions) of

$$\Psi_n(\widehat{\theta}_n) = 0 \qquad \text{or} \qquad \Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2})\,,$$

$$\Psi(\theta_0) = 0\,.$$

In the simple case of i.i.d. data $X_1, \ldots, X_n$ i.i.d. P_0 with empirical measure P_n, and then, for the usual case of linear estimating equations, the functions Ψ_n, Ψ are given by

$$\Psi_n(\theta) = \mathrm{P}_n \psi(\cdot, \theta), \qquad \text{and} \qquad \Psi(\theta) = P_0 \psi(\cdot, \theta)$$

for a vector of functions $\psi : \mathcal{X} \times \Theta \to R^k$, $\psi(x, \theta) = \underline{\psi}(x, \theta)$; often the functions ψ are score functions motivated by likelihood, pseudolikelihood, quasilikelihood, or some other "likelihood" for the data.

Here are the four basic conditions needed for Huber's theorem:

A.1

$$\sqrt{n}(\Psi_n - \Psi)(\theta_0) \to_d \mathbb{Z}_0$$

A.2

$$\sup_{|\theta - \theta_0| \le \delta_n} \frac{|\sqrt{n}(\Psi_n - \Psi)(\theta) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)|}{1 + \sqrt{n}|\theta - \theta_0|} = o_p(1)$$

for every sequence $\delta_n \to 0$.

A.3 The function Ψ is (Fréchet-)differentiable at θ_0 with nonsingular derivative $\dot{\Psi}(\theta_0) \equiv \dot{\Psi}_0$:

$$\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}_0(\theta - \theta_0) = o(|\theta - \theta_0|)\,.$$

A.4 $\Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2})$ and $\Psi(\theta_0) = 0$.

Theorem. (Huber (1967); Pollard (1985)). Suppose that A.1 - A.4 hold. Let $\widehat{\theta}_n$ be random maps into $\Theta \subset R^k$ satisfying $\widehat{\theta}_n \to_p \theta_0$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d -\dot{\Psi}_0^{-1}(\mathbb{Z}_0)\,;$$

if $\mathbb{Z}_0 \sim N_k(0, A)$, then this yields, with $\dot{\Psi}_0 \equiv B$,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d N_k(0, B^{-1}A(B^{-1})^T)\,.$$

Proof. By definition of $\widehat{\theta}_n$ and θ_0,

$$\begin{aligned}
\sqrt{n}(\Psi(\widehat{\theta}_n) - \Psi(\theta_0)) &= \sqrt{n}(\Psi(\widehat{\theta}_n) - \Psi_n(\widehat{\theta}_n)) + o_p(1) \\
&= -\sqrt{n}(\Psi_n - \Psi)(\theta_0) \\
&\quad - \left\{\sqrt{n}(\Psi_n - \Psi)(\widehat{\theta}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)\right\} + o_p(1) \\
&= -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_p(1 + \sqrt{n}|\widehat{\theta}_n - \theta_0|) + o_p(1)\,; \qquad (0.1)
\end{aligned}$$

here the last equality holds by A.2 and $\widehat{\theta}_n \to_p \theta_0$. Since $\dot{\Psi}_0$ is continuously invertible, there exists a constant $c > 0$ such that

$$\|\dot{\Psi}_0(\theta - \theta_0)\| \geq c\|\theta - \theta_0\|$$

for every θ; this is just the basic property of a nonsingular matrix. By A.3 (differentiability of Ψ), this yields

$$|\Psi(\theta) - \Psi(\theta_0)| \geq c|\theta - \theta_0| + o(|\theta - \theta_0|)\,.$$

By (0.1) it follows that

$$\sqrt{n}|\widehat{\theta}_n - \theta_0|(c + o_p(1)) \leq O_p(1) + o_p(1 + \sqrt{n}|\widehat{\theta}_n - \theta_0|)\,,$$

which implies

$$\sqrt{n}|\widehat{\theta}_n - \theta_0| = O_p(1)\,.$$

Hence from (0.1) again and A.3 it follows that

$$\dot{\Psi}_0(\sqrt{n}(\widehat{\theta}_n - \theta_0)) + o_p(\sqrt{n}|\widehat{\theta}_n - \theta_0|) = -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_p(1)$$

and therefore

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d -\dot{\Psi}_0^{-1}(\mathbb{Z}_0)$$

by A.1 and A.3. □

Now our goal is to extend this to an infinite-dimensional setting in which Θ is a Banach space. A sufficiently general Banach space is the space

$$l^{\infty}(H) \equiv \{z : H \to R \Big| \, ||z|| = \sup_{h \in H} |z(h)| < \infty\}$$

where H is a collection of functions. We suppose that

$$\Psi_n : \Theta \to L \equiv l^{\infty}(H') \, , \quad n = 1, 2, \, ...$$

are random, and that

$$\Psi : \Theta \to L \equiv l^{\infty}(H') \, ,$$

is deterministic. Suppose that either

$$\Psi_n(\widehat{\theta}_n) = 0 \qquad \text{in} \qquad L;$$

(i.e. $\Psi_n(\widehat{\theta}_n)(h') = 0$ for all $h' \in H'$), or

$$\Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2}) \qquad \text{in} \qquad L;$$

(i.e. $||\Psi_n(\widehat{\theta}_n)||_{H'} = o_p(n^{-1/2})$).
Here are the four basic conditions needed for the infinite-dimensional version of Huber's theorem due to Van der Vaart (1995):

B.1

$$\sqrt{n}(\Psi_n - \Psi)(\theta_0) \Rightarrow \mathbb{Z}_0 \qquad \text{in} \qquad l^{\infty}(H') \, .$$

B.2

$$\sup_{||\theta - \theta_0|| \le \delta_n} \frac{\|\sqrt{n}(\Psi_n - \Psi)(\theta) - \sqrt{n}(\Psi_n - \Psi)(\theta_0)\|}{1 + \sqrt{n}||\theta - \theta_0||} = o_p(1)$$

for every sequence $\delta_n \to 0$.

B.3 The function Ψ is (Fréchet-)differentiable at θ_0 with derivative $\dot{\Psi}(\theta_0) \equiv \dot{\Psi}_0$ having a bounded (continuous) inverse:

$$\|\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}_0(\theta - \theta_0)\| = o(\|\theta - \theta_0\|) \, .$$

B.4 $\Psi_n(\widehat{\theta}_n) = o_p(n^{-1/2})$ in $l^{\infty}(H')$ and $\Psi(\theta_0) = 0$ in $l^{\infty}(H')$.

Theorem. (Van der Vaart, 1995). Suppose that B.1 - B.4 hold. Let $\widehat{\theta}_n$ be random maps into $\Theta \subset l^{\infty}(H')$ satisfying $\widehat{\theta}_n \to_p \theta_0$. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \Rightarrow -\dot{\Psi}_0^{-1}(\mathbb{Z}_0) \qquad \text{in} \qquad l^{\infty}(H) \, .$$

Proof. Exactly the same as in the finite-dimensional case. □

Applications of Van der Vaart's Z-theorem:

- Gamma frailty model; Murphy (1995).
- Partially censored data; Van der Vaart (1995).
- Correlated gamma-frailty model; Parner (1998).
- Double-censoring; Chang (19xx)
- Semiparametric biased sampling models; Gilbert (1997).
- Two-phase sampling models with data missing by design; Breslow and Holubkov (1997), Lawless, Wild, and Kalfbleisch (1997), McNeney and Wellner (1998).

However, in many statistical problems the parameter usually includes both a finite-dimensional parameter (e.g. regression parameters) and an infinite dimensional (nuisance) parameter. We now suppose that $\theta = (\beta, \Lambda)$, where β is a finite-dimensional parameter, say in $\mathbb{R}^d$, and Λ an infinite dimensional parameter (a function). The M-estimators of β, $\hat{\beta}_n$, and of Λ, $\hat{\Lambda}_n$, respectively, often have different convergence rates. The convergence rate for $\hat{\Lambda}_n$ is often smaller than $n^{1/2}$, such as $n^{1/3}$, or $n^{2/5}$ in some cases. Huang (1996) established a general theorem to show that under certain hypotheses, the maximum likelihood estimator of a finite dimensional parameter has $n^{1/2}$ convergence rate and is asymptotically semiparametric efficient, even though the convergence rate for the maximum likelihood estimator of the infinite dimensional parameter is smaller than $n^{1/2}$. He also successfully applied his general theorem to the proportional hazards model with interval censored data.

The following theorem due to Zhang (1998) generalizes the theorem of Huang (1996) to the case of inefficient M-estimators; it shows that under reasonable regularity hypotheses, the M-estimator of a finite-dimensional parameter β has $n^{1/2}$ convergence rate, and that $\hat{\beta}_n$ is asymptotically normal, even though the M-estimator of the corresponding infinite dimensional parameter Λ converges perhaps more slowly than $n^{1/2}$. The resulting asymptotic covariance matrix for the M-estimator of β has the well-known "sandwich" structure.

Here is the notation and conditions needed for the theorem. Let $\theta = (\beta, \Lambda)$, where $\beta \in \mathbb{R}^d$, and Λ is an infinite dimensional parameter in a class of functions $\mathcal{F}$. Λ_η is a parametric path in $\mathcal{F}$ through Λ, i.e. $\Lambda_\eta \in \mathcal{F}$, and $\Lambda_\eta|_{\eta=0} = \Lambda$.

Let $\mathbf{H} = \left\{ h : h = \frac{\partial \Lambda_\eta}{\partial \eta}\Big|_{\eta=0} \right\}$ and define

$$m_1(\beta, \Lambda; x) = \nabla_\beta m_{(\beta,\Lambda)}(x) \equiv \left(\frac{\partial}{\partial \beta_1} m_{(\beta,\Lambda)}(x), \cdots, \frac{\partial}{\partial \beta_d} m_{(\beta,\Lambda)}(x) \right)'.$$

$$m_2(\beta,\Lambda;x)[h] = \frac{\partial}{\partial\eta} m_{(\beta,\Lambda_\eta)}(x)\bigg|_{\eta=0},$$
$$m_{11}(\beta,\Lambda;x) = \nabla^2_\beta m_{(\beta,\Lambda)}(x),$$
$$m_{12}(\beta,\Lambda;x)[h] = \frac{\partial}{\partial\eta} m_1(\beta,\Lambda_\eta;x)\bigg|_{\eta=0},$$
$$m_{21}(\beta,\Lambda;x)[h] = \nabla_\beta m_2(\beta,\Lambda;x)[h],$$

and

$$m_{22}(\beta,\Lambda;x)[h,h] = \frac{\partial^2}{\partial\eta^2} m(\beta,\Lambda_\eta;x)\bigg|_{\eta=0}.$$

We also define

$$S_1(\beta,\Lambda) = Pm_1(\beta,\Lambda;X),$$
$$S_2(\beta,\Lambda)[h] = Pm_2(\beta,\Lambda;X)[h],$$
$$S_{1n}(\beta,\Lambda) = \mathbb{P}_n m_1(\beta,\Lambda;X),$$
$$S_{2n}(\beta,\Lambda)[h] = \mathbb{P}_n m_2(\beta,\Lambda;X)[h],$$
$$\dot{S}_{11}(\beta,\Lambda) = Pm_{11}(\beta,\Lambda;X),$$
$$\dot{S}_{12}(\beta,\Lambda)[h] = \dot{S}'_{21}(\beta,\Lambda)[h] = Pm_{12}(\beta,\Lambda;X)[h],$$

and

$$\dot{S}_{22}(\beta,\Lambda)[h,h] = Pm_{22}(\beta,\Lambda;X)[h,h].$$

Furthermore, for $\mathbf{h} = (h_1,\cdots,h_d)' \in \mathbf{H}^d$, where $h_j \in \mathbf{H}$ for $j = 1,2,\cdots,d,$, and $\mathbf{H}^d = \underbrace{\mathbf{H}\times\mathbf{H}\times\cdots\times\mathbf{H}}_{d}$, denote

$$m_2(\beta,\Lambda;x)[\mathbf{h}] = (m_2(\beta,\Lambda;x)[h_1],\cdots,m_2(\beta,\Lambda;x)[h_d])',$$
$$m_{12}(\beta,\Lambda;x)[\mathbf{h}] = (m_{12}(\beta,\Lambda;x)[h_1],\cdots,m_{12}(\beta,\Lambda;x)[h_d]),$$
$$m_{21}(\beta,\Lambda;x)[\mathbf{h}] = (m_{21}(\beta,\Lambda;x)[h_1],\cdots,m_{21}(\beta,\Lambda;x)[h_d]),$$
$$m_{22}(\beta,\Lambda;x)[\mathbf{h},h](m_{22}(\beta,\Lambda;x)[h_1,h],\cdots,m_{22}(\beta,\Lambda;x)[h_d,h])^T,$$

and define

$$S_2(\beta,\Lambda)[\mathbf{h}] = Pm_2(\beta,\Lambda;X)[\mathbf{h}],$$
$$S_{2n}(\beta,\Lambda)[\mathbf{h}] = \mathbb{P}_n m_2(\beta,\Lambda;X)[\mathbf{h}],$$
$$\dot{S}_{12}(\beta,\Lambda)[\mathbf{h}] = Pm_{12}(\beta,\Lambda;X)[\mathbf{h}],$$
$$\dot{S}_{21}(\beta,\Lambda)[\mathbf{h}] = Pm_{21}(\beta,\Lambda;X)[\mathbf{h}],$$

and

$$\dot{S}_{22}(\beta,\Lambda)[\mathbf{h},h] = Pm_{22}(\beta,\Lambda;X)[\mathbf{h},h].$$

The following Assumptions will be used to formulate our general theorem:

A1. **(Consistency and rate of convergence):**

$$|\hat{\beta}_n - \beta_0| = o_p(1) \quad \text{and} \quad \|\hat{\Lambda}_n - \Lambda_0\| = O_p(n^{-\gamma})$$

for some $\gamma > 0$.

A2. **(Zero-mean structure):**

$$S_1(\beta_0, \Lambda_0) = 0, \quad \text{and} \quad S_2(\beta_0, \Lambda_0)[h] = 0, \quad \text{for all } h \in \mathbf{H}.$$

A3. **(Positive "pseudo-information"):** There exists an $\mathbf{h}^* = (h_1^*, \cdots, h_d^*)^T$, $h_j^* \in \mathbf{H}$ $j = 1, \cdots, d$, such that

$$\dot{S}_{12}(\beta_0, \Lambda_0)[h] - \dot{S}_{22}(\beta_0, \Lambda_0)[\mathbf{h}^*, h] = 0, \tag{0.2}$$

for all $h \in \mathbf{H}$. Moreover, the matrix

$$A = -\dot{S}_{11}(\beta_0, \Lambda_0) + \dot{S}_{21}(\beta_0, \Lambda_0)[\mathbf{h}^*] = -P\left(m_{11}(\beta_0, \Lambda_0; X) - m_{21}(\beta_0, \Lambda_0; X)[\mathbf{h}^*]\right)$$

is nonsingular.

A4. **(Approximate solution of pseudo-score equations):** The estimator $(\hat{\beta}_n, \hat{\Lambda}_n)$ satisfies

$$S_{1n}(\hat{\beta}_n, \hat{\Lambda}_n) = o_{p^*}(n^{-1/2}),$$

and

$$S_{2n}(\hat{\beta}_n, \hat{\Lambda}_n)[\mathbf{h}^*] = o_{p^*}(n^{-1/2}).$$

A5. **(Stochastic equicontinuity):** For any $\delta_n \downarrow 0$ and $C > 0$,

$$\sup_{|\beta-\beta_0|\le\delta_n, \|\Lambda-\Lambda_0\|\le Cn^{-\gamma}} \left|\sqrt{n}(S_{1n} - S_1)(\beta, \Lambda) - \sqrt{n}(S_{1n} - S_1)(\beta_0, \Lambda_0)\right| = o_{p^*}(1),$$

and

$$\sup_{|\beta-\beta_0|\le\delta_n, \|\Lambda-\Lambda_0\|\le Cn^{-\gamma}} \left|\sqrt{n}(S_{2n} - S_2)(\beta, \Lambda)[\mathbf{h}^*] - \sqrt{n}(S_{2n} - S_2)(\beta_0, \Lambda_0)[\mathbf{h}^*]\right| = o_{p^*}(1).$$

A6. **(Smoothness of the model):** For some $\alpha > 1$ satisfying $\alpha\gamma > 1/2$, and for (β, Λ) in the neighborhood $\{(\beta, \Lambda) : |\beta - \beta_0| \le \delta_n, \|\Lambda - \Lambda_0\| \le Cn^{-\gamma}\}$,

$$\begin{aligned}\Big|S_1(\beta, \Lambda) - S_1(\beta_0, \Lambda_0) - \dot{S}_{11}(\beta_0, \Lambda_0)(\beta - \beta_0) - \dot{S}_{12}(\beta_0, \Lambda_0)[\Lambda - \Lambda_0]\Big| \\ = o(|\beta - \beta_0|) + O(\|\Lambda - \Lambda_0\|^\alpha),\end{aligned}$$

$$\begin{aligned}\Big|S_2(\beta, \Lambda)[\mathbf{h}^*] - S_2(\beta_0, \Lambda_0)[\mathbf{h}^*] - \dot{S}_{21}(\beta_0, \Lambda_0)[\mathbf{h}^*])(\beta - \beta_0) - (\dot{S}_{22}(\beta_0, \Lambda_0)[\mathbf{h}^*, \Lambda - \Lambda_0]\Big| \\ = o(|\beta - \beta_0|) + O(\|\Lambda - \Lambda_0\|^\alpha).\end{aligned}$$

A7. **(Asymptotic normality of projected pseudo-score):** With

$$m^*(\beta_0, \Lambda_0; x) \equiv m_1(\beta_0, \Lambda_0; x) - m_2(\beta_0, \Lambda_0; x)[\mathbf{h}^*],$$

we have

$$\sqrt{n}\mathbb{P}_n m^*(\beta_0, \Lambda_0; X) \longrightarrow_d N(0, B),$$

where $B = Em^*(\beta_0, \Lambda_0; X)^{\otimes 2} = Em^*(\beta_0, \Lambda_0; X)m^*(\beta_0, \Lambda_0; X)'$.

Theorem 2.3.5. (Asymptotic Normality) Suppose that: assumptions A1-A7 hold. Then

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = A^{-1}\sqrt{n}\mathbb{P}_n m^*(\beta_0, \Lambda_0; X) + o_{p^*}(1) \longrightarrow_d N\left(0, A^{-1}B\left(A^{-1}\right)'\right).$$

Proof : A1 and A5 yield

$$\sqrt{n}(S_{1n} - S_1)(\hat{\beta}_n, \hat{\Lambda}_n) - \sqrt{n}(S_{1n} - S_1)(\beta_0, \Lambda_0) = o_{p^*}(1).$$

Since $S_{1n}(\hat{\beta}_n, \hat{\Lambda}_n) = o_{p^*}(n^{-1/2})$ by A4 and $S_1(\beta_0, \Lambda_0) = 0$ by A2, it follows that

$$\sqrt{n}S_1(\hat{\beta}_n, \hat{\Lambda}_n) + \sqrt{n}S_{1n}(\beta_0, \Lambda_0) = o_{p^*}(1).$$

Similarly, we have that

$$\sqrt{n}S_2(\hat{\beta}_n, \hat{\Lambda}_n)[\mathbf{h}^*] + \sqrt{n}S_{2n}(\beta_0, \Lambda_0)[\mathbf{h}^*] = o_{p^*}(1).$$

Combining these equalities and A6 yields

$$\begin{aligned} &\dot{S}_{11}(\beta_0, \Lambda_0)[\hat{\beta}_n - \beta_0] + \dot{S}_{12}(\beta_0, \Lambda_0)[\hat{\Lambda}_n - \Lambda_0] + S_{1n}(\beta_0, \Lambda_0) \\ &\qquad + o(|\hat{\beta}_n - \beta_0|) + O(\|\hat{\Lambda}_n - \Lambda_0\|^{\alpha}) = o_{p^*}(n^{-1/2}), \end{aligned} \tag{0.3}$$

$$\begin{aligned} &\dot{S}_{21}(\beta_0, \Lambda_0)[\mathbf{h}^*][\hat{\beta}_n - \beta_0] + \dot{S}_{22}(\beta_0, \Lambda_0)[\mathbf{h}^*][\hat{\Lambda}_n - \Lambda_0] + S_{2n}(\beta_0, \Lambda_0)[\mathbf{h}^*] \\ &\qquad + o(|\hat{\beta}_n - \beta_0|) + O(\|\hat{\Lambda}_n - \Lambda_0\|^{\alpha}) = o_{p^*}(n^{-1/2}). \end{aligned} \tag{0.4}$$

Because $\alpha\gamma > 1/2$, then the rate of convergence assumption 1 implies

$$\sqrt{n}O(\|\hat{\Lambda}_n - \Lambda_0\|^{\alpha}) = o_{p^*}(1).$$

Thus by A4 and (2.3.4) minus (2.3.5), it follows that

$$\begin{aligned} &(\dot{S}_{11}(\beta_0, \Lambda_0) - \dot{S}_{21}(\beta_0, \Lambda_0)[\mathbf{h}^*](\hat{\beta}_n - \beta_0) + o(|\hat{\beta}_n - \beta_0|) \\ &\qquad = -(S_{1n}(\beta_0, \Lambda_0) - S_{2n}(\beta_0, \Lambda_0)[\mathbf{h}^*]) + o_{p^*}(n^{-1/2}), \end{aligned}$$

i.e.

$$-(A+o(1))(\hat{\beta}_n-\beta_0) = -\mathbb{P}_n m^*(\beta_0,\Lambda_0;X) + o_{p^*}(n^{-1/2}).$$

Hence

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n-\beta_0) &= (A+o(1))^{-1}\sqrt{n}\mathbb{P}_n m^*(\beta_0,\Lambda_0;X) + o_{p^*}(1)\\ &\to_d N\left(0, A^{-1}B\left(A^{-1}\right)'\right).\end{aligned}$$

□

References:

- Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore. Softcover edition published by Springer-Verlag, to appear May (1998).

- Hu, Hui-lin. (1998). Large sample theory for Pseudo-Maximum Likelihood Estimates in Semiparametric Models. *Unpublished Ph.D. dissertation*, University of Washington, Seattle.

- Huang, J. (1996). Efficient Estimation for the Cox Model with Interval Censoring. *Annals of Statistics* **24**, 540-568.

- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1**, 221 - 233. Univ. California Press.

- Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory* **1**, 295 - 314.

- Van der Vaart, A. W. (1995a). Efficiency of infinite dimensional M-estimators. *Statistica Neerl.* **49**, 9 - 30.

- Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

- Zhang, Y. (1998). Estimation for Counting processes with Incomplete Data. Unpublished Ph.D. dissertation, University of Washington.